Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Markov Chain Marginal Bootstrap

Xuming He & Feifang Hu

To cite this article: Xuming He & Feifang Hu (2002) Markov Chain Marginal Bootstrap, Journal of the American Statistical Association, 97:459, 783-795, DOI: 10.1198/016214502388618591

To link to this article: https://doi.org/10.1198/016214502388618591

Markov Chain Marginal Bootstrap

Xuming He and Feifang Hu

Markov chain marginal bootstrap (MCMB) is a new method for constructing confidence intervals or regions for maximum likelihood estimators of certain parametric models and for a wide class of M estimators of linear regression. The MCMB method distinguishes itself from the usual bootstrap methods in two important aspects: it involves solving only one-dimensional equations for parameters of any dimension and produces a Markov chain rather than a (conditionally) independent sequence. It is designed to alleviate computational burdens often associated with bootstrap in high-dimensional problems. The validity of MCMB is established through asymptotic analyses and illustrated with empirical and simulation studies for linear regression and generalized linear models.

KEY WORDS: Asymptotic normality; Confidence interval; Generalized linear model; M estimator; Maximum likelihood; Regression.

1. INTRODUCTION

The term *bootstrap* has become a household word in statistics today, being widely known for its efficacy in bias reduction, uncertainty estimation, and confidence interval construction. From the pioneering work of Efron (1979) to the more recent writings by Hall (1992) and Efron and Tibshirani (1993), a vast amount of literature has accompanied the development of bootstrap methods over the past two decades. A primary role of bootstrap at work is the determination of reliable confidence limits when a more traditional method, such as that based on asymptotic normality, lacks efficacy. DiCiccio and Romano (1988) and DiCiccio and Efron (1996) have provided good reviews in this area.

Bootstrap methods typically provide more reliable confidence limits at the cost of computational agony. Although today's powerful computing technology has provided much of the relief, the ever-increasing size of data and complexity of models that we face argues for less computationally intensive devices. In this article we propose a new variant of bootstrap, the *Markov chain marginal bootstrap* (MCMB), that aims to provide faster computation in multiparameter M estimation problems. Our focus is on the M estimators of linear regression and the maximum likelihood estimators of other parametric models.

Given independent observations $\mathbf{Y} = \{Y_i, i = 1, \ldots, n\}$, an M estimator typically solves an estimating equation

$$S(\mathbf{Y}, \theta) = \sum_{i=1}^{n} g_i(Y_i, \theta) = 0,$$

for some score function g_i. Generalized estimating equation estimators in the statistics and biostatistics literature also fall into this category. We assume that $\theta \in R^p$ and g_i maps from the space of (y, θ) into R^p. The classical bootstrap method requires solving the p-dimensional equations for a large number of bootstrap samples from $\{Y_i\}$. Lahiri (1992) provided asymptotic properties of bootstrapping M estimators in regression. An alternative approach is to use resampling to approximate the distribution of $S = \sum_{i=1}^{n} g_i(Y_i, \theta)$ and then convert this into a confidence region for θ. This idea has been used in specific settings by Parzen, Wei, and Ying (1994) and Hu and Zidek (1995) and in more general settings by Hu and Kalbfleisch (1997, 2000). If $\theta \in R$ and $\sum_{i=1}^{n} g_i(Y_i, \theta)$ is monotone in θ, then a confidence interval of θ can be obtained from the percentiles of S^* without actually solving the estimating equation for all of the resamples. In general, however, all existing bootstrap methods rely on solving the p-dimensional systems repeatedly.

For the moment, we restrict ourselves to the linear regression problem $Y_i = x_i'\beta + e_i$, $(i = 1, \ldots, n)$, with iid errors e_i. An M estimator solves

$$n^{-1} \sum_{i=1}^{n} \psi(Y_i - x_i'\beta) x_i = 0 \tag{1}$$

for a score function ψ. We are often interested in smooth M estimators with bounded and continuous ψ functions. An important exception is the least absolute deviation (LAD) estimator with $\psi(r) = \operatorname{sgn}(r)$. In this case, (1) may not be solved exactly, but minimization of $\sum_{i=1}^{n} |Y_i - x_i'\beta|$ over $\beta \in R^p$ guarantees a solution so that (1) holds asymptotically. Treatments for the LAD estimator and generalized M estimators are discussed in more detail later in the article.

As we see from the finite-sample examples in Section 4, the confidence intervals derived from the asymptotic distribution of $\hat{\beta}_n$ can be highly variable and the confidence levels attained are sometimes too low. This is due partly to the difficulty in estimating the constant $(E\psi')^2$ needed in the asymptotic variance. Furthermore, the performance of such confidence intervals is usually highly nonrobust against violations of the iid error assumption. Bootstrap methods, although more computationally expensive, are often useful with a higher degree of stability.

To reduce the computational cost of bootstrap in the cases with modest to large p, we propose a new method to approximate the joint distribution of $n^{1/2}(\hat{\beta}_n - \beta)$. The new method generates a Markov chain for each component of β by solving p one-dimensional equations in lieu of a p-dimensional system. In the common cases with a monotone score function, finiding the solution to a one-dimensional equation is easy and numerically reliable. The variance-covariance matrix of the limiting distribution of $n^{1/2}(\hat{\beta}_n - \beta)$ can then be approximated by the second moments of the Markov chain, and confidence intervals or regions can be obtained as well.

Xuming He is Professor, Department of Statistics, University of Illinois, Champaign, IL 61820 (E-mail: *he@stat.uiuc.edu*). Feifang Hu is Assistant Professor, Department of Statistics, University of Virginia, Charlottesville, VA 22904. This research is partially supported by National Science Foundation grant DMS-0102411, University of Illinois Campus Research Board, and National University of Singapore's Research Grant. The authors thank Stephen Portnoy and Qiman Shao for helpful discussions and two anonymous referees and the associate editor for constructive comments on earlier versions of the article.

Journal of the American Statistical Association
September 2002, Vol. 97, No. 459, Theory and Methods
DOI 10.1198/016214502388618591

There appears to be some similarity between MCMB and the well-known Markov chain Monte Carlo (MCMC) methods, such as the Gibbs sampler. Both approaches aim to reduce computational complexity by breaking up a high-dimensional problem into lower-dimensional problems. However, MCMB has a unique structure, as we discuss in the following sections.

The rest of the article is organized as follows. Section 2 gives the MCMB method for estimating equations in a general parametric model with a special emphasis on linear regression estimation. Roughly speaking, an estimator is said to be MCM bootstrappable when the MCM bootstrap produces a Markov chain with the same asymptotic distribution as the estimator itself. This property is shown for M estimators of regression with smooth score functions as well as for the LAD estimator. For general parametric models, Section 3 shows that the maximum likelihood estimators (MLEs) with concave likelihood functions are generally MCM bootstrappable with the generalized linear models as a special case. Section 4 reports a Monte Carlo study for the finite-sample performance of the MCMB confidence intervals for regression and makes some comparisons with several other known methods. Robustness of the MCMB method against certain deviations from homoscedasticity is demonstrated. Savings in computer time of MCM bootstrap over the traditional bootstrap is more explicitly examined for problems of modest to large dimensions. A real data example and a Monte Carlo simulation with logistic regression are also given to illustrate how MCMB works in nonlinear models. Section 5 provides some further discussions on the proposed method and point to some possible extensions and improvements. The Appendix gives technical proofs.

2. METHOD AND THEORY

Let $Y_1, \ldots, Y_n$ be a sequence of independent random vectors and let $\theta \in \Omega \subset R^p$ be an unknown p-dimensional parameter associated with the distributions of Y_i. Suppose that $g_i(y, \theta)$ are R^p-valued functions such that $E_\theta\{g_i(Y_i, \theta)\} = 0$ for all $i = 1, \ldots, n$ and $\theta \in \Omega$. Here and in the rest of the article, E_θ denotes the expectation under the model given θ, and var_θ represents the variance-covariance under the same model. We consider $\hat{\theta}$, an estimator of θ, as a solution to the *unbiased estimating equation*

$$S(\mathbf{Y}, \theta) = \sum_{i=1}^{n} g_i(Y_i, \theta) = 0. \tag{2}$$

Such estimators are often called M estimators in the statistics literature when they come from maximizing a likelihood-type criterion, in which case (2) is just the gradient equation of the objective function. However, (2) is sometimes motivated for its own sake and is generally called an estimating equation (see Godambe and Kale 1991). We suppose that $S(\mathbf{Y}, \theta)$ is a one-to-one function of θ so that the solution to (2) is unique. We briefly discuss cases with multiple roots in Section 5.

2.1 The Algorithm

Let $\hat{\theta}$ be an estimator of $\theta = (\theta_1, \ldots, \theta_p)'$. To present the algorithm for a general context, assume that we have a decomposition $g_i(Y_i, \theta) = a_i z_i$, where the a_i's are constant vectors or matrices independent of $\mathbf{Y}$ and the z_i's are random variables depending on $\mathbf{Y}$. If a_i is a vector, then z_i is scaler. If a_i is a scaler or a matrix, then z_i is a vector here. If $a_i \neq a_j$ for some (i, j), then we need z_i to be (asymptotically) exchangeable. This decomposition of g_i is obviously not unique, thus allowing different schemes for resampling. One simple scheme is to set $a_i = 1$ for all i; then no exchangeability requirement on z_i is necessary to approximate the distribution of $S(\mathbf{Y}, \theta)$ through resampling. The conventional bootstrap solves (2) repeatedly for randomly generated data from the empirical distribution of Y_i or from the estimating equation $\sum_i g_i(Y_i, \hat{\theta})$. The MCM bootstrap that we propose follows a few simple steps, as follows. (We examine the validity of the method later.)

1. Initialize $\hat{\theta}^{(0)} = \hat{\theta}$ and $k = 1$.
2. Draw a bootstrap sample $\{z_{1j}^{*(k)}, \ldots, z_{nj}^{*(k)}\}$ from $\{z_1, \ldots, z_n\}$ for each $j = 1, \ldots, p$.
3. In the sequence of $j = 1, 2, \ldots, p$, solve for $\hat{\theta}_j^{(k)}$ from

$$S_j(\mathbf{Y}, \hat{\theta}_1^{(k)}, \ldots, \hat{\theta}_{j-1}^{(k)}, \hat{\theta}_j^{(k)}, \hat{\theta}_{j+1}^{(k-1)}, \ldots, \hat{\theta}_p^{(k-1)}) = S_j^{*(k)}, \tag{3}$$

   where S_j is the jth component of $S(\mathbf{Y}, \theta)$, and $S_j^{*(k)}$ is the jth component of $\sum_{i=1}^{n} a_i z_{ij}^{*(k)}$.
4. Increase k by 1 and go to step 2, or stop if k has reached a prespecified level.

From this process, we get an MCMB sequence $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \ldots$. It is clear by construction that this is a Markov chain, not a set of independent realizations. We hope that the bootstrap distribution of $n^{1/2}(\hat{\theta} - \hat{\theta}^{(k)})$ approximates the sampling distribution of $n^{1/2}(\hat{\theta} - \theta)$ when both n and k are large. Furthermore, if the sequence is nearly ergodic, then approximate confidence intervals or regions for θ can be constructed from just one chain.

The MCM bootstrap is reminiscent of the Gibbs sampler, which approximates the joint distribution through conditionals. Both share some of the highly desirable properties of MCMC methods. For example, MCMB reduces the problem of solving a p-dimensional system to p one-dimensional equations, making it possible to handle high-dimensional problems. However, it does not really fit into the standard MCMC framework. It does not work with the conditional distribution of $\hat{\theta}^*$. An important feature of MCMB is that at the kth iteration, we do not sample from the p-dimensional vector g_i; rather, we sample (independently) from each component of g_i, and as a result, the correlation among the components of g_i is not kept. The validity of the MCMB algorithm depends on whether the correlation structure can be recovered through the marginal updating part of the algorithm.

Although not universally applicable in the present form, the MCMB method is shown to be asymptotically consistent for M estimators of linear regression and for the MLEs of a rather general class of parametric models. In the next section we formulate the asymptotic validity of MCMB proposed here.

2.2 Asymptotic Validity

Suppose that an estimator $\hat{\theta}$ is asymptotically normal such that $n^{1/2}(\hat{\theta} - \theta) \to N(0, \Sigma)$. In the cases we consider in this article, we show that for some choice of (a_i, z_i) the Markov chain $\hat{\theta}^{(k)}$ obtained from the MCMB algorithm of Section 2.1

is approximately a multiple AR(1) process in the following sense.

For each n, there exists a stationary ergodic sequence $Z_{n,k}$ ($k = \ldots, -1, 0, 1, \ldots$) satisfying a p-dimensional autoregressive model

$$Z_{n,k} = \Lambda_n Z_{n,k-1} + u_{n,k} \tag{4}$$

with the following properties:

(P1) Λ_n is a p by p matrix converging to Λ whose absolute eigenvalues are all less than 1.

(P2) For each n, $u_{n,k}$ ($|k| = 0, 1, 2, \ldots$) are i.i.d random vectors with mean 0 and finite second moments.

(P3) The distribution of $u_{n,1}$ converges to $N(0, H)$ as $n \to \infty$ with $H = \Sigma - \Lambda\Sigma\Lambda'$.

(P4) $E_*\|n^{1/2}(\hat{\theta} - \hat{\theta}^{(k)}) - Z_{n,k}\|^2 \to 0$ almost surely as $n, k \to \infty$ such that $k \le K_n$, where K_n is some sequence tending to infinity with n.

Throughout the article, E_* and P_* denote expected value and probability under the bootstrap distribution conditional on the sample. In (P4), the almost sure convergence is in the original probability space of $\mathbf{Y}$.

If the properties (P1)–(P4) hold, then we say that the estimator is *MCM bootstrappable*. A necessary condition for the estimator $\hat{\theta}$ to be MCM bootstrappable is that

$$P_*(\sqrt{n}(\hat{\theta} - \hat{\theta}^{(k)}) \le t) \to P(N(0, \Sigma) \le t) \tag{5}$$

in probability for all $t \in \mathcal{R}^p$. Note that the left side of (5) is a function of the data $\mathbf{Y}$ and thus is a random quantity (see Shao and Tu 1995, pp. 72–79 for more details). Property (P4) implies that for large n and k, the bootstrap sequence of $n^{1/2}(\hat{\theta} - \hat{\theta}^{(k)})$ behaves like the AR(1) sequence $Z_{n,k}$. The following lemma gives a weak ergodic property of $Z_{n,k}$. (For more details on stationary ergodic processes, see Hanann 1970.)

Lemma 1. Under the properties (P1)–(P3) for the multiple AR(1) processes $Z_{n,k}$ in (4) and as $n, m \to \infty$, we have (a) $m^{-1}\sum_{k=1}^{m} Z_{n,k} \to 0$ in probability, (b) $Z_{n,m}$ converges in distribution to $N(0, \Sigma)$, and (c) $m^{-1}\sum_{k=1}^{m} Z_{n,k}Z_{n,k}' \to \Sigma$ in probability if $\sup_n E\|u_{n,k}\|^{2+\epsilon} < \infty$ for some $\epsilon > 0$.

It is then clear that we can construct approximate confidence intervals or regions by using a single Markov chain $\hat{\theta}^{(k)}$. In fact, property (P4) and Lemma 1 imply that under the bootstrap distribution, $m^{-1}\sum_{k=1}^{m} \hat{\theta}^{(k)} - \hat{\theta} \to 0$ and $(n/m)\sum_{k=1}^{m}(\hat{\theta}^{(k)} - \hat{\theta})(\hat{\theta}^{(k)} - \hat{\theta})' \to \Sigma$ in probability provided that $m, n \to \infty$ and $m \le K_n$. In our empirical studies in Section 4, confidence intervals are constructed using values from single chains.

Some additional comments are in order.

Remark 1. The asymptotic theory for MCMB involves both the sample size n and the length of the Markov chain K_n going to infinity. We must restrict ourselves to the cases where K_n increases with n at a controlled rate (such as a power of n). This is to avoid accumulation of error through iteration. In this article we do not address the question of how large K_n can be relative to n, but see Section 5 for further comments.

Remark 2. In general, $g_i(Y_i, \hat{\theta})$ depends on the parameter estimate $\hat{\theta}$, so a loss of p degrees of freedom is expected. As is common with most approximation methods, the finite-sample performance is typically enhanced if $S_j^{*(k)}$ is replaced by $\sqrt{n/(n-p)}S_j^{*(k)}$ in (3). In some special cases, the distribution of $g_i(Y_i, \theta)$ is known even without knowing θ. Then we can sample directly from this distribution.

2.3 Application to Linear Models

Consider the following multiple linear regression model:

$$Y_i = x_i'\beta + e_i, \qquad i = 1, \ldots, n, \tag{6}$$

where $e_1, \ldots, e_n$ are independent random variables with the common probability density function f, $x_1, \ldots, x_n$ are given design vectors, and β is a p-dimensional parameter to be estimated. Throughout this article, A' denotes the transpose of a vector or matrix A. The M estimator $\hat{\beta}_n$ of β corresponding to ψ is defined as a solution to the vector equation

$$\sum_{i=1}^{n} x_i\psi(Y_i - x_i'\beta) = 0, \tag{7}$$

where ψ is a real-valued function satisfying $E\psi(e_i) = 0$ and $\mathrm{var}(\psi(e_i)) = \sigma^2$ for $i = 1, \ldots, n$. The MCMB algorithm can be carried out with $a_i = x_i$ and $z_i = \psi(y_i - x_i'\hat{\beta}_n)$. Let

$$\mathbf{Q}_n = n^{-1}\sum_{i=1}^{n} x_i x_i'$$

and

$$\gamma = \int \psi'(e) f(e)\, de = -\int \psi(e) f'(e)\, de,$$

where ψ' and f' denote the derivatives of ψ and f. The distinction between the transpose of a matrix and the derivative of a function should be clear from the context. Also note that our definition of γ requires that only one of the foregoing integrals exist.

In this section we show that M estimators of linear regression are MCM bootstrappable. We state formal conditions later, but the basic idea is best illustrated by the least squares estimator with $p = 2$ for simplicity. In this special case, we have

$$n^{-1/2}\sum_{i=1}^{n}\left(Y_i - x_{i1}\hat{\beta}_1^{(k)} - x_{i2}\hat{\beta}_2^{(k-1)}\right)x_{i1} = d_1^{(k)}$$

and

$$n^{-1/2}\sum_{i=1}^{n}\left(Y_i - x_{i1}\hat{\beta}_1^{(k)} - x_{i2}\hat{\beta}_2^{(k)}\right)x_{i2} = d_2^{(k)}, \tag{8}$$

where $d_1^{(k)} = n^{-1/2}\sum_{i=1}^{n} x_{i1}e_{i1}^{*(k)}$ and $d_2^{(k)} = n^{-1/2}\sum_{i=1}^{n} x_{i2}e_{i2}^{*(k)}$, and both $e_{i1}^{*(k)}$ and $e_{i2}^{*(k)}$ are independent bootstrap samples from the empirical distribution of $r_i = Y_i - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}$ ($i = 1, \ldots, n$), the residuals from the parameter estimate (β_1, β_2). Also, let $s_{11} = n^{-1}\sum_{i=1}^{n} x_{i1}^2$, $s_{12} = n^{-1}\sum_{i=1}^{n} x_{i1}x_{i2}$, and $s_{22} = n^{-1}\sum_{i=1}^{n} x_{i2}^2$. Then the two equations can be written as

$$s_{11}n^{1/2}\left(\hat{\beta}_1 - \hat{\beta}_1^{(k)}\right) = d_1^{(k)} - s_{12}n^{1/2}\left(\hat{\beta}_2 - \hat{\beta}_2^{(k-1)}\right)$$

and

$$s_{22}n^{1/2}\big(\hat\beta_2-\hat\beta_2^{(k)}\big)=d_2^{(k)}-s_{12}n^{1/2}\big(\hat\beta_1-\hat\beta_1^{(k)}\big).$$

Note that under the bootstrap distribution, the right sides of these equations are sums of two independent variables. Thus, by taking variance-covariance operation and assuming that the covariance matrix of $n^{1/2}(\hat\beta-\hat\beta^{(k)})$ stabilizes to $V=(v_{ij})_{2\times2}$ as $k\to\infty$, we have

$$s_{11}^2v_{11}=s_{11}\sigma^2+s_{12}^2v_{22},$$
$$s_{22}^2v_{22}=s_{22}\sigma^2+s_{12}^2v_{11},$$

and

$$s_{22}v_{12}=-s_{12}v_{11}.$$

Direct calculations show that $\mathbf{V}=\sigma^2((s_{ij})_{2\times2})^{-1}$. That is, the bootstrap variance-covariance of $n^{1/2}(\hat\beta-\hat\beta^{(k)})$ stabilizes to the desired asymptotic covariance matrix for the least squares estimator.

Remark 3. For the M estimator defined in (7), we often choose $a_i=x_i$ and $z_i=\psi(r_i)-\bar\psi$ with $\bar\psi=n^{-1}\sum_{i=1}^n\psi(r_i)$. The centering of $\psi(r_i)$ is a finite-sample adjustment to improve performance. Another possible scheme is to choose $a_i=1$ and $z_i=x_i\psi(r_i)$.

2.3.1 M Estimators With Smooth Score Functions. We first consider the simple case where the estimator minimizes $\sum_{i=1}^n\rho(y_i-x_i'\beta)$ for some convex objective function ρ with the score function $\psi=\rho'$. It is then well known that

$$(n\mathbf{Q}_n)^{1/2}(\hat\beta-\beta)\to N(0,(\sigma/\gamma)^2\mathbf{I}) \tag{9}$$

under mild conditions on $\mathbf{Q}_n$ (see Yohai and Maronna 1979; He and Shao 1996). The normal approximation is accurate up to the order of $n^{-1/2}$.

For any univariate function f, we say that it is Lipschitz if there exists a constant $C(f)<\infty$ such that $|f(x)-f(y)|\le C(f)|x-y|$ for all x,y. Let $r_i=y_i-x_i'\hat\beta$ and $z_i=\psi(r_i)-\bar\psi$ as in Remark 3. The MCM bootstrap proceeds with resampling from the z_i's. The asymptotic property of the MCMB sequence $\hat\beta^{(k)}$ is given as follows.

Theorem 1. If the following conditions (C1)–(C3) are satisfied, then any M estimator satisfying (7) is MCM bootstrappable:

(C1) ψ' is Lipschitz.

(C2) $E\psi(e)=0$, $E|\psi(e)|^3<\infty$, and $\gamma=E\{\psi'(e)\}\ne0$.

(C3) (a) $\mathbf{Q}_n\to\mathbf{Q}$ for some positive definite matrix $\mathbf{Q}$, and $\sum_{i=1}^n\|x_i\|^3=O(n)$ as $n\to\infty$, and (b) The average $n^{-1}\sum_{i=1}^n x_i\psi(y_i-x_i'\beta)$ converges as $n\to\infty$ to a continuously differentiable function $U(\beta)$ as a gradient of some strictly convex function of β.

Note that $U(\beta)$ may depend on the given design sequence x_i, but if x_i is a random sample from some distribution with finite third moment and ψ is continuously differentiable, then the condition (C3) is automatically satisfied with probability 1 and $U(\beta)=E(x_1\psi(y_1-x_1'\beta))$. When f is smooth, (C3) can hold even when ρ is not strictly convex and not differentiable everywhere.

By Theorem 1, we have

$$P(n^{1/2}(\hat\beta-\beta)\le x)-P_*(n^{1/2}(\hat\beta-\hat\beta^{(k)})\le x)\to0 \tag{10}$$

in probability as both $k(\le K_n)$ and n tend to infinity.

2.3.2 Minimum L_q Distance Estimators. We now consider the minimum L_q distance estimators defined through minimization of $\sum_{i=1}^n|Y_i-x_i'\beta|^q$ for some $1<q<2$. The cases of higher q values are covered in the previous section. Asymptotic normality results for these estimators have been given by Bai, Rao, and Wu (1992), and the accuracy of first-order approximations was given by He and Shao (1996). Under (C3) and the following two conditions:

(C4) $\sum_{i=1}^n E(|e-x_i'\beta|^q-|e|^q)$ has a unique minimum at $\beta=0$, and

(C5) f is bounded and $E|e-t|^{q-2}$ is Lipschitz in t,

we have

$$n^{1/2}(\hat\beta-\beta)\to N\big(0,\sigma_q^2\mathbf{Q}^{-1}\big) \tag{11}$$

with

$$\sigma_q^2=(E|e|^{2q-2})\{(q-1)E|e|^{q-2}\}^{-2}.$$

The following theorem establishes the validity of the MCM bootstrap for these estimators.

Theorem 2. For $1<q<2$ and under conditions (C3)–(C5), the minimum L_q distance estimator is MCM bootstrappable.

Construction of confidence intervals based on the asymptotic variance-covariance of the L_q-norm estimator requires estimating $E|e|^{q-2}$. A natural estimate is $n^{-1}\sum_{i=1}^n|r_i|^{q-2}$, but for $q<2$, this estimate tends to have a large variability. Our experience shows that the accuracy of such confidence intervals is often poor as a result (see the simulation results in Sec. 4.2). The MCM bootstrap avoids direct estimation of this quantity, thus resulting in more stable and more accurate confidence intervals or variance estimates.

2.3.3 The Least Absolute Deviation Regression. The minimum L_1 distance estimator is of special interest in that the distribution of $\psi(e_i)$ is known and independent of the parameter β when the error e_i has zero median. In this case, we can sample directly from a binomial distribution for $\sum z_i$. This is similar to the resampling method of Parzen et al. (1994), except that we use independent components as described in Section 2.1. To be more specific, (3) is replaced by

$$\sum_{i=1}^n x_{ij}\operatorname{sgn}\left(Y_i-\sum_{l=1}^{j-1}x_{il}\hat\beta_l^{(k)}-x_{ij}\beta_j^{(k)}-\sum_{l=j+1}^{p}x_{il}\hat\beta_l^{(k-1)}\right)=\sum_{i=1}^n x_{ij}\big(2b_{ij}^{(k)}-1\big) \tag{12}$$

for $j=1,\dots,p$ and $k=1,2,\dots$, where $b_{ij}^{(k)}$ are iid Bernoulli random variables with success probability of $1/2$.

Because $\psi(r)=\operatorname{sgn}(r)$ is a jump function, (12) may not be satisfied exactly. But the left side of (12) is monotone in $\beta_j^{(k)}$, so we can define the root as the point at which

the left side crosses the right side. The method of Parzen et al. (1994) solves vector equations. These authors resolved the problem of jump discontinuity by converting the equation solving into an L_1 minimization problem with a pseudo-observation. Asymptotically, this results in a difference of only $o_P(n^{-1/2})$. The MCM bootstrap is computationally simpler for large p. The following theorem establishes its asymptotic validity. As in Section 2.3.1, we assume that the average $n^{-1}\sum_{i=1}^n x_i \operatorname{sgn}(y_i - x_i'\beta)$ converges to $U(\beta)$, the gradient of a strictly convex function of β with continuous second-order derivative matrix. This is a mild restriction on the design points x_i and holds under general conditions when (x_i, y_i) constitutes a random sample.

Theorem 3. Suppose that e_i has zero median and that its density function is Lipschitz at 0 with $f(0) > 0$. In addition to condition (C3), we assume that $\max_{1\le i\le n} |x_i| = O(n^{1/4})$. Then

$$n^{1/2}(\hat{\beta} - \beta) \to, N(0, (4f^2(0)\mathbf{Q})^{-1}), \tag{13}$$

and the estimator is MCM bootstrappable.

Construction of confidence intervals based on the asymptotic variance of LAD requires nonparametric estimation of $1/f(0)$. Alternative methods that bypass direct estimation of the sparsity include the rank inversion method of Gutenbrunner, Jurečková, Koenker, and Portnoy (1993) and the local regression quantile method of Zhou and Portnoy (1996). Koenker (1994) showed the rank inversion method to be accurate but to require rather expensive parametric programming, especially when a joint confidence region of β is needed. The local regression quantile method gives a joint confidence region for β, but its marginal projections do not lead to right confidence intervals. Moreover, it is based on the difference between two neighboring quantiles, and thus the construction is often troubled by the problem of quantile crossing with a finite sample. The MCMB method does not suffer from any of these problems. We give a simulation comparison in Section 4.1.

2.3.4 Generalized M Estimators. Generalized M (GM) estimators are robust estimators that bound the local influence of leverage points. They solve

$$\sum_{i=1}^n \psi(Y_i - x_i'\beta) x_i w(x_i) = 0 \tag{14}$$

for some weight function w such that $xw(x)$ is bounded (see Hampel, Ronchetti, Rousseeuw, and Stahel 1986 for more details). Under similar conditions for smooth M estimators with $\mathbf{Q}_n$ replaced by $n^{-1}\sum_{i=1}^n w(x_i)x_i x_i'$, the GM estimators are MCM bootstrappable. We skip the details here.

2.4 Sensitivity to Heteroscedasticity

The M estimators of linear regression are shown to be MCM bootstrappable under the model (6) with iid errors. Sensitivity of the procedure to unequal variances is an issue of practical interest. Unlike the traditional paired bootstrap that samples from the (x_i, y_i) pairs, MCMB fails to be asymptotically correct in general heteroscedastic models unless a correct likelihood is specified. However, as we show in the simulation study in the next section, it is not very sensitive to certain types of deviations from the iid errors.

Adjustments to the MCMB algorithm to suit more general heteroscedastic models are still under investigation, but we give a simple characterization for understanding the sensitivity of the MCMB method in cases with unequal variances of e_i. Let $\operatorname{var}(e_i) = \sigma_i^2$, which may depend on x_i.

For the sake of simplicity, refer to (8) again in the case where $p = 2$. Also, assume centering and scaling of the design x_i so that $\sum_i x_i = 0$ and $\mathbf{Q}_n = \mathbf{I}$. Under this reparameterization, it is easy to see that the marginal variances from MCMB are $v_{11} = \sum_i x_{i1}^2\sigma_i^2$ and $v_{22} = \sum_i x_{i1}^2\sigma_i^2$, which are the same as the marginal variances of the sampling distribution for the estimator. But the covariance term from MCMB becomes $v_{12} = \sum_i x_{i1}x_{x2}\sigma_i^2$, compared to 0 under the sampling distribution. Therefore, it is the size of $\sum_i x_{i1}x_{x2}\sigma_i^2$ that determines the sensitivity of MCMB.

If σ_i is linear in x_i and x_i is a random sample from any distribution symmetric about its mean with finite third moments, then $\sum_i x_{i1}x_{x2}\sigma_i^2$ is approximately 0 for large n. In such cases, the MCM bootstrap remains valid. This is probably not all that surprising. (For a similar robustness property of bootstrap procedures under some non-iid models, see Liu 1988.)

In general, if the design is such that the two matrices $\sum_i x_i x_i'$ and $\sum_i \sigma_i^2 x_i x_i'$ can be diagonalized simultaneously, then the MCM bootstrap reproduces the asymptotic variance-covariance structure of the estimator. The deviation is determined by the size of the off-diagonal elements of the correlation matrix for $\sigma_i \mathbf{Q}_n^{-1/2} x_i$ $(i = 1, \ldots, n)$. Examples can be found where they are close to 1, so the MCMB confidence intervals are way off, but in many situations that we have considered (including the example used in Sec. 4.1), they are close to 0, suggesting insensitivity of MCMB to certain types of heteroscedasticity. We hope to report further research in a separate work.

3. MARKOV CHAIN MARGINAL BOOTSTRAP FOR MAXIMUM LIKELIHOOD ESTIMATORS

The validity of MCM bootstrap holds for a wide variety of M estimators in linear regression. For general parametric models, we show that the same is true for MLEs when the likelihood is concave.

Suppose that Y_i $(i = 1, 2, \ldots, n)$ has the probability density function $f_i(, \theta)$ with an unknown parameter $\theta \in R^p$. The MLE $\hat{\theta}$ is obtained by maximizing $\sum_{i=1}^n \log f_i(Y_i, \theta)$. Assuming differentiability of the densities f_i, the estimator solves

$$\sum_{i=1}^n h_i(Y_i, \theta) = 0, \tag{15}$$

where $h_i(y, \theta)$ is the derivative of $\log f_i(y, \theta)$ with respect of θ. The MCM bootstrap algorithm can be carried out with $a_i = 1$ and $z_i = h_i(Y_i, \hat{\theta})$.

Theorem 4. If the following conditions (C6)–(C8) are satisfied with probability 1, then the MLE (15) is MCM bootstrappable:

(C6) As $n \to \infty$, $l_n(\theta) = n^{-1}\sum_{i=1}^n \log f_i(Y_i, \theta)$ converges to a strictly concave function of θ with a continuous second-order derivative, and $\partial l_n(\theta)/\partial\theta$ also converges.

(C7) $n^{-1}\sum_{i=1}^{n}\partial^2 h_i(Y_i,\theta)/\partial\theta^2$ is uniformly bounded in n, and $\sum_{i=1}^{n}E|h_i(Y_i,\theta)|^3 = O(n)$.

(C8) Both $n^{-1}\sum_{i=1}^{n}[h_i(Y_i,\theta)h_i(Y_i,\theta)']$ and $-n^{-1}\times\sum_{i=1}^{n}\partial h_i(Y_i,\theta)/\partial\theta$ converge to the information matrix $D(\theta) > 0$ as $n \to \infty$.

The MLEs are used in a wide variety of applications. Theorem 4 extends the applicability of the MCMB methodology. If Y_i are iid, and thus $h_i = h$ for all i, then condition (C8) follows from existence of the Fisher information matrix $E_\theta h(Y,\theta)h(Y,\theta)' = -E_\theta(\partial h(Y,\theta)/\partial\theta)$. We present an application of Theorem 4 to generalized linear models later. In this case, h_i depends on covariates x_i.

Following McCullagh and Nelder (1989), we consider the generalized linear model for Y_i that comes from an exponential family of distributions with

$$f_Y(y;\theta,\phi) = \exp\{(y\theta - b(\theta))/a(\phi) + c(y,\phi)\} \tag{16}$$

for some specific functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot,\cdot)$, and

$$EY_i = \mu_i = g(x_i'\beta), \qquad i = 1,\ldots,n, \tag{17}$$

where g is a link function, x_i are covariates, and $\beta \in R^p$ is the parameter of interest. Let $\eta_i = x_i'\beta$. For simplicity, we consider such models with the canonical link function so that $\theta = \eta = x'\beta$. Common examples include the following:

- Normal model: $\eta = \mu$ and $b(\theta) = \theta^2/2$
- Poisson model: $\eta = \log\mu$ and $b(\theta) = e^\theta$
- Logistic model: $\eta = \log(\mu/(1-\mu))$ and $b(\theta) = \log(1+e^\theta)$
- Gamma model: $\eta = 1/\mu$ and $b(\theta) = -\log(-\theta)$
- Inverse Gaussian: $\eta = 1/\mu^2$ and $b(\theta) = -(-2\theta)^{1/2}$.

We first assume that the dispersion parameter ϕ is known. We also use $b^{(l)}$ to denote the lth derivative of b.

Corollary 1. Given the dispersion parameter ϕ and the canonical link function in the generalized linear model (16) and (17), the following conditions (C9)–(C10) are sufficient for the MLE of β to be MCM bootstrappable:

(C9) $n^{-1}\sum_{i=1}^{n}x_i b^{(1)}(x_i'\beta)$ converges to a function of β as $n \to \infty$ and $n^{-1}\sum_{i=1}^{n}\|x_i\|^3 b^{(3)}(x_i'\beta)$ is bounded uniformly in n.

(C10) $n^{-1}\sum_{i=1}^{n}\|x_i\|^{2+2\delta}E|(Y_i - b^{(1)}(x_i'\beta))^2 - b^{(2)}(x_i'\beta)\times a(\phi)|^{1+\delta}$ is bounded uniformly in n for some constant $\delta > 0$, and $n^{-1}\sum_{i=1}^{n}x_i x_i' b^{(2)}(x_i'\beta) \to D(\beta)$ for some positive definite matrix $D(\beta)$.

Conditions (C9) and (C10) are easy to check for those generalized linear models discussed earlier. If x_i is a random sample from a p-variate probability distribution, then all we need is certain moment or tail conditions. Taking the logistic model for example, direct calculations show that $b^{(2)}(\theta) = e^\theta(1+e^\theta)^{-2}$ and $b^{(3)}(\theta) = e^\theta(1-e^\theta)(1+e^\theta)^{-3}$. It is clear that $|b^{(2)}(\theta)| \le 1$ and $|b^{(3)}(\theta)| < 1$; thus conditions (C9) and (C10) with $\delta = 1/2$ follow from the finite third moment of x_i.

Remark 4. When the dispersion parameter ϕ is unknown, we may plug an estimate, $\hat\phi$, into the likelihood estimating function of β. As long as ϕ is replaced by its consistent estimator $\hat\phi$ that is asymptotically orthogonal to $\hat\beta$, the results of Corollary 1 remain true.

Because $E\partial^2\log f_Y(y;\theta,\phi)/\partial\theta\partial\phi = E(Y - b^{(1)}(\theta))a^{(1)}(\phi)/a^2(\phi) = 0$, the ML estimate of β and ϕ are asymptotically orthogonal. Therefore, in the MCMB algorithm, to construct the confidence intervals for β, we can simply use the MLE of ϕ throughout the algorithm.

4. EMPIRICAL INVESTIGATION

In this section we report on a finite-sample simulation study for confidence intervals of the minimum L_q distance estimators in regression with $q = 1$ and $q = 1.5$. Comparisons with some existing methods are made. Bootstrap methods often give more reliable confidence intervals than normal approximations for such estimators. We also apply MCMB to logistic regression with a real data example and a Monte Carlo evaluation. We compare the computational complexity of MCMB and traditional bootstrap methods at the end of the section.

4.1 Least Absolute Deviation Regression

Obtaining standard errors and confidence intervals of the LAD regression estimator has attracted interest in recent years. Koenker (1994) reviewed and compared several methods. To facilitate comparison with those methods, we used the same setup here.

In the first case, samples of size $n = 50$ are generated from

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i$$

with x_{1i}, x_{2i}, x_{3i}, and e_i all independent and with the t distribution with ν degrees of freedom. The true parameters β_i are set at 0. The choices of $\nu = 3$ and 8 are used in our simulation study. We focus on 90% confidence intervals of the three slope parameters. In the simulation, we took 500 random samples and for each sample used a Markov chain of length 1000 for the MCMB method.

To obtain the confidence intervals based on the asymptotic variances of the LAD estimate, we need to estimate the sparsity $1/f(0)$. The performance of the confidence intervals depends quite heavily on this nonparametric estimate. A kernel method of Hall and Sheather (1988) is explicitly designed for this problem and, as reported by Koenker (1994), seems to perform well. The coverage probability and length of the confidence intervals (averaged over the three slope parameters) for the MCMB method are given in Table 1. The corresponding results from four other methods are taken from Koenker (1994) for comparison: RANK, the rank inversion method; NORM, normality based interval with the Hall–Sheather estimate of sparsity; PWY, the bootstrap method of Parzen et al. (1994); and PAIR, the pairwise xy bootstrap method. The bootstrap size is kept at 1000 across the board. The lower and upper 5th percentiles of the bootstrap sample are used to obtain confidence intervals. In this case with iid errors, all of the bootstrap methods tend to be conservative, and no marked differences in performance are noted. The normality based method with the chosen sparsity estimate yields very good results, as does the rank-inversion method.

Table 1. LAD Regression: Simulated Coverage Probabilities (C) and Lengths (L) of 90% Confidence Intervals Under iid Errors

		MCMB	NORM	PWY	PAIR	RANK
$\nu = 3$	C	.947	.922	.957	.948	.893
	L	.528	.455	.520	.486	.427
$\nu = 8$	C	.941	.915	.957	.945	.879
	L	.615	.613	.680	.640	.558

A more difficult problem is estimation of the confidence intervals when there is deviation from the iid errors. We now generate x_{1i}, x_{2i}, and x_{3i} as independent lognormal variates and draw e_i from a Student t distribution with ν degrees of freedom times a scale of $\sigma_i = (1 + x_{1i} + x_{2i} + x_{3i})/5$. The results in this heteroscedastic case are given in Table 2.

In this case, the method based on the asymptotic variance estimates is clearly not working; it is very sensitive to the presence of heteroscedasticity in the data. The MCMB method and the rank inversion method are not sensitive to this type of heteroscedasticity. In fact, judging from both coverage and length, the MCMB method gives the best performance in this case, which is consistent with our discussion in Section 2.5.

Note that the performance of the MCMB method relative to other methods especially designed for the LAD estimate varies from model to model. However, our experience shows that the MCM bootstrap is overall very competitive with other bootstrap methods that require more intensive computations in higher dimensional problems.

4.2 Minimum L_q Distance Estimator With $q = 1.5$

Consider the linear model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, \qquad i = 1, \ldots, n, \tag{18}$$

with x_{1i}, x_{2i}, and e_i all generated independently from the standard normal distribution. With $n = 20$ and the true parameters β_i set at 0, we consider 90% confidence intervals of all three parameters. We use both fixed designs and random designs here. The paired bootstrap (PAIR), MCMB, and the normal approximation (NORM) using the variance estimate of (11) are compared based on 500 samples drawn from the model (18). A Markov chain of length 500 is used for MCMB, and a bootstrap sample of size 500 is used for PAIR. As with Section 4.1, simple percentile confidence intervals are used from the bootstrap samples. For the fixed design, the residual-based bootstrap (RESI) is also included (see Freedman 1981).

For the fixed design case, we set $\mathbf{x}_1 = (1.27, -1.10, 2.19, .73, -.07, .42, .37, .45, -.78, .76, .44, 1.32, -.40, .33, -.40, .55, .51, -.11, -1.15, 1.71)$, and $\mathbf{x}_2 = (1.60, 1.09, -.02, -.83, 3.05, .34, -.87, .45, -.78, .76, .44, 1.32, -.40, .33, -1.85, .69, .11, 1.47, .87, .12)$, and Y_i are generated from (18). The average confidence intervals in Table 3 are obtained by taking the averages of the two endpoints of the intervals over 500 cases.

Table 3. Minimum L_q Distance Estimator ($q = 1.5$): Simulated Coverage Probabilities and Average 90% Confidence Intervals in [ ] (fixed design)

	β_0		β_1		β_2	
MCMB	91.0	[−.43, .42]	88.4	[−.45, .47]	87.8	[−.33, .33]
NORM	76.0	[−.36, .33]	76.0	[−.36, .38]	74.8	[−.28, .29]
PAIR	87.8	[−.46, .43]	86.2	[−.46, .51]	86.0	[−.37, .38]
RESI	87.8	[−.42, .40]	87.2	[−.43, .45]	87.8	[−.33, .34]

For each of the 500 samples, there is an estimate of β_0, β_1, and β_2. The standard deviations for those estimates are .25, .28, and .21. Confidence intervals constructed from these using the standard formula of average plus (minus) 1.64 times SD are [−.43, .40], [−.45, .48], and [−.34, .34]. They may be used as benchmarks for the other methods under consideration. It is clear from Table 3 that MCMB performed well.

We next generated data from (18) using random designs. Again, we used a total of 500 samples. The results are summarized in Table 4. In this case, the results for the two slope parameters are averaged. The 500 parameter estimates have standard deviations of .24 for the intercept and .25 for the slope. They suggest confidence intervals of length .80 and .82. Table 4 shows that MCMB performed best in this case.

It is interesting to note that the method of NORM resulted in a very low coverage probability. This is a result of the high variability in estimating $E|e|^{-.5}$ needed for its asymptotic variance. Averaging over $|r_i|^{-.5}$ is problematic when some of the residuals may be very close to 0. The bootstrap methods avoid this pitfall.

It is also clear that MCMB performs better than the paired bootstrap in both Tables 3 and 4. The average lengths of the confidence intervals for the slope parameters are noticeably larger with PAIR. This problem persists in our simulation studies for other minimum L_q distance estimators with q between 1 and 2. This may be explained in part by the fact that in the bootstrap samples under PAIR, not all of the design points must be selected. MCMB appears to be more stable, because there is never a change in the design matrix in the computations.

Table 2. LAD Regression: Simulated Coverage Probabilities (C) and Lengths (L) of 90% Confidence Intervals Under Heteroscedasticity

		MCMB	NORM	PWY	PAIR	RANK
$\nu = 3$	C	.909	.717	.950	.911	.902
	L	.651	.552	.907	.799	.793
$\nu = 8$	C	.909	.656	.946	.911	.902
	L	.580	.357	.715	.612	.621

Table 4. Minimum L_q Distance Estimator ($q = 1.5$): Simulated Coverage Probabilities (C) and Average 90% Confidence Intervals in [ ] (random design)

	β_0		β_1 and β_2	
MCMB	90.4	[−.40, .42]	89.2	[−.41, .43]
NORM	76.0	[−.32, .33]	77.5	[−.34, .35]
PAIR	88.0	[−.42, .44]	87.2	[−.44, .47]

4.3 Maximum Likelihood Estimates in Logistic Regression

First, we consider a real data example with the logistic model. We then use this example to examine the consistency of the MCMB method for the MLEs under nonlinear models.

The data were collected from a study of risk factors associated with low infant birth weight at Baystate Medical Center, Springfield, Massachusetts during 1986, and can be found in appendix 1 of Hosmer and Lemeshow (1989). The indicator of low birth rate is fitted by logistic regression on the following eight covariates:

AGE, age of mothers in years
LWT, weight in pounds at the last menstrual period
*RACE*1, race indicator (1 for white, 0 otherwise)
*RACE*2, race indicator (1 for black, 0 otherwise)
SMOKE, smoking indicator during pregnancy (1 for yes, 0 for no)
PTL, history of premature labor (0 for none, 1 for once, etc.)
HT, history of hypertension (1 for yes, 0 for no)
UI, Presence of uterine irritability (1 for yes, 0 for no).

With a total of 189 cases in the dataset, we used MCMB with a chain of length 1000 to estimate the variance-covariance matrix of the nine-parameter estimate (including the intercept). Using 10% level of significance, three covariates (*AGE*, *PTL*, *UI*) turned out to be (conditionally) insignificant. This agrees with the large-sample inference based on the normal approximation. Table 5 gives the estimated standard errors of all of the parameter estimates under three different methods: the asymptotic variance approach (NORM), the paired bootstrap (PAIR), and MCMB.

We note in this example that the variance estimates from MCMB are slightly higher than those from NORM, and the variances of PAIR are even higher. Our Monte Carlo experiment shows that this is quite typical. For example, we consider the logistic regression model

$$\log(P\{Y=1|x\}/(1-P\{Y=1|x\}))=\alpha+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3$$

with $(\alpha,\beta_1,\beta_2,\beta_3) = (1,-.5,1,-.5)$. The covariate x_1 is drawn from $N(0,1)$, x_2 is Bernoulli with success probability of .5, and x_3 is drawn from the lognormal distribution. We used samples of size $n = 100$ and obtained 90% confidence intervals for the three slope parameters based on the variance estimates from MCMB, PAIR, and NORM. Table 6 gives for the coverage and length of those intervals. The NORM intervals are the shortest and the PAIR intervals the longest, with the MCMB intervals in between. The coverage probabilities are all around the nominal level of 90%, and the bootstrap methods tend to be conservative. In this example, the "true" variances obtained from 2000 Monte Carlo samples are .073, .265, and .038 for the slope estimates. The average variance estimates from NORM are .067, .255, and .036. These are only slightly below the true values, probably reflecting the difference between the conditional (given x) and the unconditional variances. The average variance estimates from MCMB are .081, .330 and .047; they tend to overestimate. The averages from PAIR are even higher at .085, .345, and .051.

Table 5. Infant Birth Weight Example: Parameter Estimates (est) and Standard Errors (s.e.) From Different Methods

	est	*MCMB (s.e.)*	*NORM (s.e.)*	*PAIR (s.e.)*
INTERCEPT	1.326	1.187	1.105	1.228
AGE	−.027	.041	.036	.037
LWT	−.015	.007	.007	.008
*RACE*1	−.862	.461	.439	.503
*RACE*2	.402	.567	.539	.581
SMOKE	.923	.422	.401	.440
PTL	.542	.470	.346	.484
HT	1.834	.741	.692	1.020
UI	.7586	.528	.459	.554

Table 6. Simulated Coverage Probabilities (C) and Lengths (L) of Confidence Intervals in a Logistic Model

		NORM	*PAIR*	*BCa*	*MCMB*	*MCMBCa*
β_1	C	.904	.936	.936	.924	.902
	L	.863	.970	.935	.940	.892
β_2	C	.884	.940	.890	.930	.910
	L	1.667	1.909	1.805	1.885	1.764
β_3	C	.886	.940	.900	.914	.920
	L	.619	.724	.642	.690	.616

It seems clear that in the logistic model, normal approximations work a little better than simple bootstrap confidence intervals. We hasten to add, however, that better bootstrap confidence intervals are available. For example, the BCa method (see Efron and Tibshirani 1996, sec. 14.3) can improve both the coverage and the length of the bootstrap confidence intervals. Table 6 includes the results for the BCa method based on paired bootstrap samples (labeled "BCa") and on MCMB sequence (labeled "MCMBCa"). In all of the bootstrap calculations for Table 6, the size of resamples is 500. We note that the BCa intervals are derived for the usual bootstrap sequence, so we cannot claim higher-order accuracy for the MCMBCa intervals used here. But the empirical evidence suggests that the bias correction and acceleration methods are worth exploring for MCMB. For further discussion on MCMBCa, see Section 5.

4.4 Computational Complexity for Problems of Large Size

We now examine the amount of computation needed to do MCMB and compare this with that of the traditional bootstrap method. The cases of large n and p are of interest here, where the computational savings can be quite substantial.

The bootstrap methods consist of two stages of operation. The first stage is simple random sampling. Sampling from a given set of data with replacement requires only $O(n)$ arithmetic operations. So in this stage, the MCMB method is more costly, because it has to draw p samples instead of just one sample. The second stage is to obtain a parameter estimate. For MCMB, only $O(n)$ arithmetic operations are needed to solve for the root of each one-dimensional equation (3), so the total number of operations is in the order of $O(np)$. For the traditional bootstrap, this takes at least np^2 operations

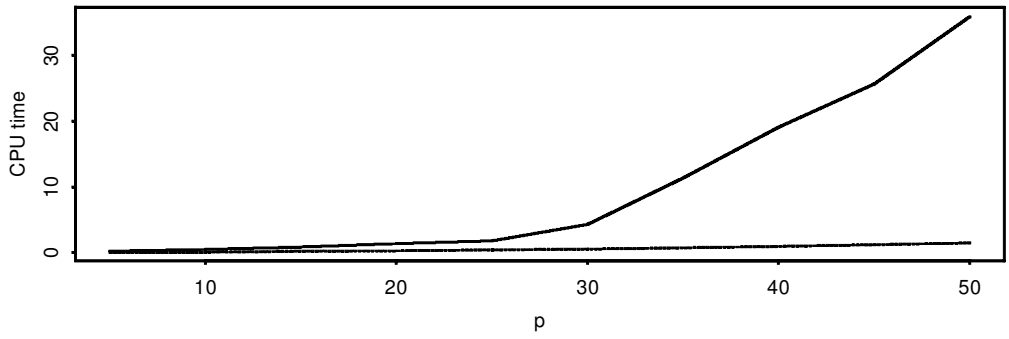

Figure 1. Timing Comparison for LAD (····· MCM bootstrap; —— paired bootstrap).

(and usually ones of higher order). Note that solving one *linear* system of p-dimensions already has a complexity of this order.

With these two stages combined, we see that the complexity of traditional bootstrap is bounded from below by Knp^2, but that of MCMB is only on the order of Knp, where K is the bootstrap sample size. Therefore, we can expect greater savings in computer time from MCMB when the dimension p is larger.

To see how the computing times compare in modest size problems, we consider a linear regression problem with p varying from 5 to 50. First, the test is on LAD ith $n = 5000$. We use *l1fit* in S-PLUS, which calls for a Fortran subroutine. We also code MCMB in Fortran with an S-PLUS interface. In the comparison, we ignore the fact that the built-in S-PLUS function *l1fit* is more efficient than our own code for MCMB. The calculations were made on a Sun Microsystems Ultra 1. For each p, a multivariate normal random sample (x_i, y_i) of size 5000 was generated, to which a bootstrap method was applied. The Unix CPU times needed to obtain parameter estimates per bootstrap sample for these two methods are plotted in Figure 1. It is clear that the effect of dimension on computing time is evident once p reaches 25 in this setting.

We also examined the performance of the MCMB method for modestly large n and p problems. When (x, y) is p-variate standard normal with $n = 5000$ and $p = 20$, we computed the MCMB chains of length $K_n = 200$. Based on a total of 200 samples, the average standard error estimate for all the slope parameters was .0174, compared to the true asymptotic value of .0177. Averaged over all of the slope parameters, the 90% confidence intervals obtained from MCMB had a coverage probability of 89.4% and average length of .0566. They compared quite favorably with the 91.4% and .0597 from the paired bootstrap method. At this time, such a simulation comparison still takes several hours to complete, and empirical studies for larger real data problems have not been made.

Our second test computes the minimum L_q distance estimator of regression with $q = 1.5$. A random sample of size $n = 1000$ is generated from multivariate normal distribution for the experiment. Newton iterations are used to solve all equations, and convergence is declared when two consecutive solutions differ by only 10^{-5}. The computations were carried out on an HP Workstation using C functions. The Unix CPU times per bootstrap sample are plotted in Figure 2.

5. DISCUSSION AND CONCLUSIONS

MCMB is a new method of resampling for constructing confidence intervals or regions with M estimators. In this article we have focused on M estimators of linear regression and the MLEs of other parametric models. We have demonstrated the applicability and merits of MCMB in a number of ways. The performance of MCMB confidence intervals are highly competitive with those of other bootstrap methods, but the reduction in computational complexity makes it possible to perform bootstrap-based inference for larger parametric models than is currently feasible with traditional bootstrap methods.

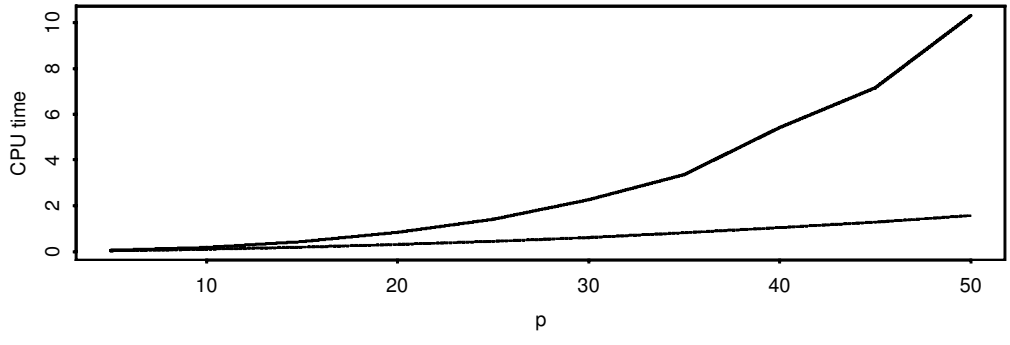

Figure 2. Timing Comparison for Minimum 1.5-Norm Estimator (····· MCM bootstrap; —— paired bootstrap).

We conclude the article with the following remarks to reflect the current status of research on MCMB:

1. Compared with the traditional resampling methods, MCMB has advantages that are probably more evident with estimators of high-dimensional parameters that are computationally expensive. We avoid the need to compute the estimates for each bootstrap sample; the only repeated operation is solving one-dimensional equations. In contrast, the MCMB algorithm given in Section 2.1 does not work for all M estimators of nonlinear models. We are working to find appropriate adjustments to broaden MCMB's applicability.

2. MCMB's asymptotic validity depends on both the chain length K_n and the sample size n tending to infinity. This makes MCMB different from both the traditional bootstrap (where K_n can be arbitrarily large) and the MCMC approach to approximate the bootstrap distribution for $\hat{\theta}^*$ (for each fixed n).

The MCMC approach would have the advantages of simpler asymptotics while still reducing dimensionality in computation. However, the distribution of $\hat{\theta}^*$ is specified only through (3), and we do not yet know any way to implement an MCMC algorithm.

In this article we have not addressed the general question of how large K_n can be relative to n. Lemma A.3 (in the Appendix) requires that K_n not grow faster than some power of n, but for bounded scores we can show that the result holds even when K_n is any power of n. Practically, the theory for the allowable size of K_n may not be so important. Although the chain length may look large for small sample problems, we prefer that it grow slowly with n. If a large amount of resampling is desired, then we suggest the alternative of running several independent chains of modest length.

3. It is often possible to make the traditional bootstrap methods for confidence intervals accurate up to the second order. The simple percentile confidence intervals from MCMB are not second-order accurate, but, as shown in Table 6, adaptation of the BCa method can provide noticeable improvement. The BCa confidence intervals from the traditional bootstrap methods are second-order accurate, but it remains an open question whether and how an adjustment to the calculation of the acceleration parameter a must be made to achieve the same for MCMBCa.

4. In this article we have considered only M estimators derived from a convex objective function. In this case the marginal equations that must be solved at each iteration of the MCMB algorithm have unique roots. For nonconvex objectives, we may have multiple roots. We believe that MCMB continues to be asymptotically valid under rather mild regularity conditions if at each iteration we choose the root closest to the estimator at the original sample. A rigorous treatment is yet to be given.

5. Because the MCMB sequence is approximately the AR(1) process of (4), numerical instability will occur when some eigenvalue of Λ_n is close to 1. The resulting MCMB sequence may also have a high autocorrelation. We are currently addressing this potential problem through reparameterization, and the details will appear elsewhere.

We anticipate that this article will open the door to a more computationally efficient approach to bootstrapping in problems with high-dimensional parameters. We believe that as its applicability extends to other types of estimators and as more of its properties are discovered, the MCMB method will fulfill an even greater promise than this article might have suggested.

APPENDIX: PROOFS OF MAIN RESULTS

We start with the proof for Lemma 1, which uses rather standard techniques in probability theory.

Proof of Lemma 1

Without loss of generality, assume that all the absolute eigenvalues of Λ_n are less than $\rho < 1$. Because $Z_{n,k} = \sum_{i=1}^{\infty} \Lambda_n^i u_{n,k-i}$, we have

$$\sum_{k=1}^{m} Z_{n,k} = \sum_{i=1}^{\infty} \Lambda_n^i \left(\sum_{k=1}^{m} u_{n,k-i} \right).$$

We have

$$E \left\| m^{-1} \sum_{k=1}^{m} Z_{n,k} \right\|^2 \le \sum_{i=1}^{\infty} \rho^{2i} E \left\| m^{-1} \sum_{k=1}^{m} u_{n,k-i} \right\|^2 \le C_1 m^{-1}$$

for some constant C_1 and uniformly for all n. By the Markov inequality, we obtain

$$P \left(\left\| m^{-1} \sum_{k=1}^{m} Z_{n,k} \right\| > \delta \right) \le C_2 \delta^2 / m$$

for any $\delta > 0$ and some constant C_2. Conclusion (a) of the lemma then follows.

To prove (c), use

$$\begin{aligned} Z_{n,k} Z'_{n,k} &= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \Lambda_n^i u_{n,k-i} u'_{n,k-j} \left(\Lambda_n^j \right)' \\ &= \sum_{i=1}^{\infty} \Lambda_n^i u_{n,k-i} u'_{n,k-i} \left(\Lambda_n^i \right)' + \sum_{i \ne j} \Lambda_n^i u_{n,k-i} u'_{n,k-j} \left(\Lambda_n^j \right)' \end{aligned}$$

and

$$m^{-1} \sum_{k=1}^{m} Z_{n,k} Z'_{n,k} - \Sigma = m^{-1} \sum_{k=1}^{m} Z_{n,k} Z'_{n,k} - \Sigma_n + (\Sigma_n - \Sigma),$$

where Σ_n is the covariance matrix of $Z_{n,k}$. Because the application of the Markov inequality can be evoked with any $(1+\epsilon)$th moment, the same technique used for (a) shows that $m^{-1} \sum_{k=1}^{m} Z_{n,k} Z'_{n,k} - \Sigma_n \to 0$ in probability. With (P3), it is easy to see that $\Sigma_n \to \Sigma$ and thus the conclusion (c) of the lemma follows.

To prove (b), we calculate the characteristic function of $Z_{n,k}$ and find its limit to be that of $N(0, \Sigma)$ as $n \to \infty$. The calculations are straightforward and can be obtained from the authors.

The following result on eigenvalues is important for the proofs of our main theorems in Sections 2 and 3. First, for any matrix $\mathbf{T}$, let $\mathcal{L}(\mathbf{T})$ denote the matrix of same size, but all of the upper off-diagonal elements of $\mathbf{T}$ are replaced by 0. That is, if t_{ij} $(1 \le i, j \le p)$ are the ijth element of $\mathbf{T}$, then write

$$\mathbf{A} = \mathcal{L}(\mathbf{T}) = \begin{pmatrix} t_{11} & 0 & \cdots & 0 \\ t_{21} & t_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ t_{p1} & t_{p2} & \cdots & t_{pp} \end{pmatrix}$$

and

$$\mathbf{B} = \mathbf{T} - \mathbf{A} = \begin{pmatrix} 0 & t_{12} & \cdots & t_{1p} \\ 0 & 0 & \cdots & t_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}.$$

We also use diag($\mathbf{T}$) as the matrix by taking only the diagonal elements of $\mathbf{T}$.

Lemma A.1. For any $p \times p$ real positive definite matrix $\mathbf{T}$, let $\mathbf{A} = \mathcal{L}(\mathbf{T})$ and $\mathbf{B} = \mathbf{T} - \mathbf{A}$. Then the absolute eigenvalues of $\mathbf{A}^{-1}\mathbf{B}$ are all less than 1.

Proof. We show that for any $\lambda = \lambda_1 + i\lambda_2$ satisfying $|\lambda| \le 1$, the determinant of $\mathbf{A} - \lambda\mathbf{B}$ is nonzero. If $\lambda = -1$, then $\mathbf{A} - \lambda\mathbf{B} = \mathbf{T}$, and the result is trivial. Thus we consider only the case with $\lambda \ne -1$ but $|\lambda| \le 1$. We show that for any $p \times 1$ real vector $\mathbf{v}_1$ and $\mathbf{v}_2$ with $\mathbf{v} = \mathbf{v}_1 + i\mathbf{v}_2 \ne 0$,

$$\bar{\mathbf{v}}'(\mathbf{A} - \lambda\mathbf{B})\mathbf{v} \ne 0.$$

Here $\bar{\mathbf{v}}' = \mathbf{v}_1' - i\mathbf{v}_2'$ is the conjugate of $\mathbf{v}$.

Direct calculations show that

$$\begin{aligned}\bar{\mathbf{v}}'(\mathbf{A} - \lambda\mathbf{B})\mathbf{v} = {} & \frac{1-\lambda_1}{2}\bar{\mathbf{v}}'\mathbf{T}\mathbf{v} + \frac{1+\lambda_1}{2}\bar{\mathbf{v}}'\mathbf{T}_1\mathbf{v} - \bar{\mathbf{v}}'[i(\lambda_2\mathbf{B} - \lambda_2\mathbf{B}')/2]\mathbf{v} \\ & + i\{-\bar{\mathbf{v}}'[(1+\lambda_1)(\mathbf{B} - \mathbf{B}')/(2i)]\mathbf{v} \\ & - \bar{\mathbf{v}}'[\lambda_2(\mathbf{B} + \mathbf{B}')/2]\mathbf{v}\}, \end{aligned} \tag{A.1}$$

where $\mathbf{T}_1 = \text{diag}(\mathbf{T})$.

If the imaginary part of $\bar{\mathbf{v}}'(\mathbf{A} - \lambda\mathbf{B})\mathbf{v}$ is equal to 0, then

$$-\bar{\mathbf{v}}'[(1+\lambda_1)(\mathbf{B} - \mathbf{B}')/(2i)]\mathbf{v} - \bar{\mathbf{v}}'[\lambda_2(\mathbf{B} + \mathbf{B}')/2]\mathbf{v} = 0,$$

which implies that

$$\bar{\mathbf{v}}'(\mathbf{B} - \mathbf{B}')\mathbf{v} = -\frac{\lambda_2}{1+\lambda_1}\bar{\mathbf{v}}'(\mathbf{B} + \mathbf{B}')\mathbf{v}i,$$

so the real part of $\bar{\mathbf{v}}'(\mathbf{A} - \lambda\mathbf{B})\mathbf{v}$ is equal to

$$\begin{aligned}\mathbf{v}_1'\left(\frac{1-\lambda_1}{2}\mathbf{T} + \frac{1+\lambda_1}{2}\mathbf{T}_1\right)\mathbf{v}_1 + \mathbf{v}_2'\left(\frac{1-\lambda_1}{2}\mathbf{T} + \frac{1+\lambda_1}{2}\mathbf{T}_1\right)\mathbf{v}_2 \\ - \frac{\lambda_2^2}{2(1+\lambda_1)}(\mathbf{v}_1'(\mathbf{B} + \mathbf{B}')\mathbf{v}_1 + \mathbf{v}_2'(\mathbf{B} + \mathbf{B}')\mathbf{v}_2)\end{aligned}$$

or, equivalently,

$$\mathbf{v}_1' D\mathbf{v}_1 + \mathbf{v}_2' D\mathbf{v}_2,$$

with

$$\begin{aligned} D &= \frac{1-\lambda_1}{2}\mathbf{T} + \frac{1+\lambda_1}{2}\mathbf{T}_1 - \frac{\lambda_2^2}{2(1+\lambda_1)}(\mathbf{B} + \mathbf{B}') \\ &= \frac{1-\lambda_1^2-\lambda_2^2}{2(1+\lambda_1)}\mathbf{T} + \frac{(1+\lambda_1)^2+\lambda_2^2}{2(1+\lambda_1)}\mathbf{T}_1. \end{aligned}$$

When $|\lambda| \le 1$ and $\lambda \ne -1$, we have

$$\frac{1-\lambda_1^2-\lambda_2^2}{2(1+\lambda_1)} > 0 \qquad \text{and} \qquad \frac{(1+\lambda_1)^2+\lambda_2^2}{2(1+\lambda_1)} > 0.$$

This means that the real part of $\bar{\mathbf{v}}'(\mathbf{A} - \lambda\mathbf{B})\mathbf{v}$ cannot be 0 whenever its imaginary part is.

To make the proofs of our theorems more focused and easier to follow, we first consider the case of least squares regression. The proof in this case is indicative of the multivariate autoregressive structure associated with MCMB. Proofs for the other cases essentially show that the same structure holds with certain error bounds.

Lemma A.2. For the least squares regression with $\mathbf{Q}_n \to \mathbf{Q}$ for some positive definite matrix $\mathbf{Q}$, the estimator is MCM bootstrappable with

$$P_*((nQ_n)^{1/2}(\hat{\beta} - \hat{\beta}^{(m)}) \le t) \to P(N(0, \sigma^2 I) \le t), \tag{A.2}$$

in probability for all $t \in \mathcal{R}^p$, when both $m, n \to \infty$.

Proof. It is easy to see that $\hat{\beta}^{(m)}$ satisfies the process

$$\mathbf{A}_n(\hat{\beta} - \hat{\beta}^{(m)}) = -\mathbf{B}_n(\hat{\beta} - \hat{\beta}^{(m-1)}) + n^{-1/2}\mathbf{e}_m^*, \tag{A.3}$$

where $\mathbf{A}_n = \mathcal{L}(\mathbf{Q}_n)$, $\mathbf{B}_n = \mathbf{Q}_n - \mathbf{A}_n$, and

$$\mathbf{e}_m^* = n^{-1/2}\left(\sum_{i=1}^n x_{i1}e_{i1}^{*(m)}, \sum_{i=1}^n x_{i2}e_{i2}^{*(m)}, \ldots, \sum_{i=1}^n x_{ip}e_{ip}^{*(m)}\right)', \tag{A.4}$$

with $e_{ij}^{*(m)}$, $i = 1, \ldots, n$ for each j, being iid random variables following the empirical distribution of centralized $\{z_i : i = 1, \ldots, n\}$, where $z_i = Y_i - x_i'\hat{\beta}$.

Rewriting (A.3), we get

$$n^{1/2}(\hat{\beta} - \hat{\beta}^{(m)}) = -\mathbf{A}_n^{-1}\mathbf{B}_n n^{1/2}(\hat{\beta} - \hat{\beta}^{(m-1)}) + \mathbf{A}_n^{-1}\mathbf{e}_m^*. \tag{A.5}$$

By Lemma A.1, we know that $\mathbf{A}_n^{-1}\mathbf{B}_n$ has all eigenvalues inside a unit sphere for sufficiently large n. Without loss of generality, we assume in the proof that this is true for all n.

Next, we define a new multiple AR(1) process as

$$\mathbf{Z}_{n,t} = -\mathbf{A}_n^{-1}\mathbf{B}_n\mathbf{Z}_{n,t-1} + \mathbf{A}_n^{-1}\mathbf{e}_t^*, \tag{A.6}$$

where $\mathbf{Z}_{n,t} \in R^p$ and $\mathbf{e}_t^*$ are iid random vectors as defined in (A.4) for $t = \ldots, -2, -1, 0, 1, 2, \ldots$. The dependence of Z_t on n is suppressed in the notation. From a well-known property of an AR(1) process (cf Hannan 1970), (A.6) defines a stationary and ergodic process. The variance-covariance matrix $\mathbf{M}_n$ of this stationary process must satisfy

$$\mathbf{M}_n = \mathbf{A}_n^{-1}\mathbf{B}_n\mathbf{M}_n\left(\mathbf{A}_n^{-1}\mathbf{B}_n\right)' + \hat{\sigma}^2\mathbf{A}_n^{-1}\,\text{diag}(\mathbf{Q}_n)\left(\mathbf{A}_n^{-1}\right)', \tag{A.7}$$

where diag($\mathbf{Q}_n$) is the matrix by taking only the diagonal elements of $\mathbf{Q}_n$, and $\hat{\sigma}^2$ is the variance of $e_{ij}^{*(m)}$. Clearly, $\hat{\sigma}^2$ converges to σ^2.

It is straightforward to check that $\mathbf{M}_n = \mathbf{Q}_n^{-1}\hat{\sigma}^2$ is a solution to (A.7). We also show that the solution to (A.7) is unique. To do so, we note that it suffices to show that the determinant of

$$\left(\mathbf{I}_{p^2 \times p^2} - \left(\mathbf{A}_n^{-1}\mathbf{B}_n\right) \otimes \left(\mathbf{A}_n^{-1}\mathbf{B}_n\right)'\right)$$

is not 0, where $(\mathbf{A}_n^{-1}\mathbf{B}_n) \otimes (\mathbf{A}_n^{-1}\mathbf{B}_n)'$ is the Kronecker product of $(\mathbf{A}_n^{-1}\mathbf{B}_n)$ and $(\mathbf{A}_n^{-1}\mathbf{B}_n)'$. This follows from the fact that the eigenvalues of $(\mathbf{A}_n^{-1}\mathbf{B}_n) \otimes (\mathbf{A}_n^{-1}\mathbf{B}_n)'$ are products of the eigenvalues of $(\mathbf{A}_n^{-1}\mathbf{B}_n)$ and $(\mathbf{A}_n^{-1}\mathbf{B}_n)'$ (see Searle 1982) and that both are in the unit sphere. Therefore, $\mathbf{M}_n = \mathbf{Q}_n^{-1}\hat{\sigma}^2$ is the variance-covariance matrix of $\mathbf{Z}_{n,t}$.

In addition, by Lemma 1, we have $\mathbf{Z}_{n,t}$ converging in distribution to $N(0, \sigma^2\mathbf{Q}^{-1})$ as $n \to \infty$.

Now compare the two sequences $\mathbf{Z}_{n,m}$ and $\hat{\beta} - \hat{\beta}^{(m)}$ to get

$$\begin{aligned} n^{1/2}(\hat{\beta} - \hat{\beta}^{(m)}) - \mathbf{Z}_{n,m} &= (-\mathbf{A}_n^{-1}\mathbf{B}_n)^m(n^{1/2}(\hat{\beta} - \hat{\beta}^{(0)}) - \mathbf{Z}_{n,0}) \\ &= -(-\mathbf{A}_n^{-1}\mathbf{B}_n)^m\mathbf{Z}_{n,0}, \end{aligned} \tag{A.8}$$

from which the conclusion of the lemma follows.

The following lemma is useful for uniform bounds on the error term in the proof of Theorem 1. It follows from Bernstein's inequality and standard truncation techniques (see also the inequality of Petrov 1955, p. 78, prob. 2.6.5).

Lemma A.3. Let $e_{n,i}$ be iid random variables with mean 0 and finite third moment, and let $\sum_{i=1}^n \|x_i\|^3 = O(n)$. Let $e_{n,i}^{(k)}$ $(k = 1, 2, \ldots)$ be independent copies of $e_{n,i}$. Then we have $\max_{k \le K_n} |n^{-1}\sum_{i=1}^n x_i e_{n,i}^{(k)}| = o(1)$ as $n \to \infty$ and $K_n = o(n^{1-\epsilon})$ for any $\epsilon > 0$.

One consequence of Lemma A.3 is that under the conditions of Theorem 1, there exists $K_n \to \infty$ such that the Markov chain $\hat{\beta}^{(k)}$ converges to β uniformly for $k \le K_n$ as $n \to \infty$. This can be shown by convexity of the objective function and uniform convergence of $n^{-1}\sum_{i=1}^n x_i\psi(Y_i - x_i'\beta)$ and its derivative to their expected values over any compact set of $\beta \in R^p$.

Proof of Theorem 1

For simplicity of presentation, we consider only the case of $p = 2$, the same arguments are valid for general cases. The sequence $\hat{\beta}^{(k)}$ is generated from a pair of equations,

$$\sum_{i=1}^n x_{i1}\psi\big(Y_i - x_{i1}\hat{\beta}_1^{(k)} - x_{i2}\hat{\beta}_2^{(k-1)}\big) = n^{1/2}s_{n,k,1} \tag{A.9}$$

and

$$\sum_{i=1}^n x_{i2}\psi\big(Y_i - x_{i1}\hat{\beta}_1^{(k)} - x_{i2}\hat{\beta}_2^{(k)}\big) = n^{1/2}s_{n,k,2}, \tag{A.10}$$

where $s_{n,k,j} = n^{-1/2}\sum_{i=1}^n x_{ij}e_{ij}^{*(k)}$, $j = 1, 2$, and $e_{ij}^{*(k)}$ are independent bootstrapped samples from $\{z_i\}$. Note that $Es_{n,k,j}^2$ is uniformly bounded in n, k, and j for $k \le K_n$. Also, we have $\max_{k\le K_n}\|\beta^{(k)} - \beta\| \to 0$ as $n \to \infty$.

Now using the mean value theorem to (A.9), we have

$$n^{-1}\sum_i x_{i1}^2\psi'(r_{k,1})n^{1/2}\big(\hat{\beta}_1^{(k)} - \hat{\beta}_1\big) + n^{-1}\sum_i x_{i1}x_{i2}\psi'(r_{k,1})n^{1/2}\big(\hat{\beta}_2^{(k-1)} - \hat{\beta}_2\big) = -s_{n,k,1} \tag{A.11}$$

where $|r_{k,1} - (Y_i - x_i'\hat{\beta})| \le |x_{i1}(\hat{\beta}_1^{(k)} - \hat{\beta}_1)| + |x_{i2}(\hat{\beta}_2^{(k-1)} - \hat{\beta}_2)|$.

Because ψ' is Lipschitz, we have

$$n^{-1}\sum_i x_{i1}^2\psi'(r_{k,1}) - n^{-1}\sum_i x_{i1}^2\psi'(Y_i - x_i'\hat{\beta}) = o\left(n^{-1}\sum_i |x_{i1}|^3\right) = o(1)$$

uniformly in $k \le K_n$. On the other hand, $n^{-1}\sum_i x_{i1}^2\psi'(Y_i - x_i'\hat{\beta}) \to q_{11}\gamma$ with q_{ij} being the ijth element of $\mathbf{Q}$. Therefore, $n^{-1}\sum_i x_{i1}^2\psi'(r_{k,1}) \to q_{11}\gamma$. Similarly, $n^{-1}\sum_i x_{i1}x_{i2}\psi'(r_{k,1}) \to q_{12}\gamma$ uniformly in $k \le K_n$ as $n \to \infty$.

Expanding (A.10), we get

$$n^{-1}\sum_i x_{i1}x_{i2}\psi'(r_{k,2})n^{1/2}\big(\hat{\beta}_1^{(k)} - \hat{\beta}_1\big) + n^{-1}\sum_i x_{i2}^2\psi'(r_{k,2})n^{1/2}\big(\hat{\beta}_2^{(k)} - \hat{\beta}_2\big) = -s_{n,k,2} \tag{A.12}$$

for some $r_{k,2}$ similar to $r_{k,1}$, and $n^{-1}\sum_i x_{i1}x_{i2}\psi'(r_{k,2}) \to q_{12}\gamma$, $n^{-1}\sum_i x_{i2}^2\psi'(r_{k,2}) \to q_{22}\gamma$ uniformly in $k \le K_n$ as $n \to \infty$.

Combining (A.11) with (A.12), we get an iterative equation

$$n^{1/2}\big(\hat{\beta}_1^{(k)} - \hat{\beta}_1\big) = r_{n,k}n^{1/2}\big(\hat{\beta}_1^{(k-1)} - \hat{\beta}_1\big) + d_{n,k}$$

where $r_{n,k} \to q_{12}^2/(q_{11}q_{22}) < 1$ uniformly in $k \le K_n$ and $d_{n,k}$ is a linear combination of $s_{n,k,j}$ $(j = 1, 2)$ with coefficients bounded uniformly in $k \le K_n$ for sufficiently large n. This means that $n^{1/2}(\hat{\beta}_1^{(k)} - \hat{\beta}_1) = \sum_{i=1}^k r_{n,k}^i d_{n,k-i}$, where $Ed_{n,k}^2$ is bounded uniformly in $k \le K_n$. As a result, $n^{1/2}(\hat{\beta}_1^{(k)} - \hat{\beta}_1)$ has a uniformly bounded second moment.

In general, the same arguments show that $\max_{k\le K_n} E\|\hat{\beta} - \hat{\beta}^{(k)}\|^2 = O(1/n)$. Using this, together with the Lipschitz assumption on ψ', we obtain

$$n^{1/2}(\hat{\beta}^{(k)} - \hat{\beta}) = -\mathbf{A}_n^{-1}\mathbf{B}_n n^{1/2}(\hat{\beta}^{(k-1)} - \hat{\beta}) + \mathbf{A}_n^{-1}(\mathbf{e}_k^*/\gamma) + R_{n,k},$$

where $\max_{k\le K_n} ER_{n,k}^2 \to 0$.

Let $Z_{n,k}$ be the multiple AR(1) series as defined in the proof for Lemma A.2. We have

$$n^{1/2}(\hat{\beta}^{(k)} - \hat{\beta}) - \gamma^{-1}Z_{n,k} = \sum_{i=1}^k\big(-\mathbf{A}_n^{-1}\mathbf{B}_n\big)^i R_{n,k-i}, \tag{A.13}$$

whose second moment goes to 0 as $k \le K_n$ and $n \to \infty$. The proof is then complete.

Proof of Theorems 2 and 3

We sketch the proof only for the case where $q = 1$.

We follow the same lines as in the proof of Theorem 2.1, but to use the mean value theorem for expansion, we replace $\sum_i x_i\psi(Y_i - x_i't)$ by its expected value. It suffices to show that the resulted difference is uniformly small. To be precise, let

$$\Delta_n(t) = \sup_{\|a-\beta\|,\|b-\beta\|\le t}\Big|\sum_i x_i\big\{\operatorname{sgn}(Y_i - x_i'a) - \operatorname{sgn}(Y_i - x_i'b) - E\operatorname{sgn}(Y_i - x_i'a) + E\operatorname{sgn}(Y_i - x_i'b)\big\}\Big|.$$

It follows from Bai and He (1999, lem. 3.2) that for any bounded $t_n > 0$,

$$P(n^{-1/2}\Delta_n(t_n) > s) \le A\exp\{-s^2/t_n\} \tag{A.14}$$

for some constant $A < \infty$. Because $E(X) = \int_0^\infty P(X \ge s)\,ds$ for any nonnegative random variable X, we conclude from (A.14) that $\max_{k\le K_n} E\Delta_n^2(t_k) \to 0$ for any bounded sequence t_k as $n \to \infty$.

The proof of Theorem 1 can then be duplicated with $s_{n,k,j}$ replaced by $s_{n,k,j}$ plus an independent error term whose second moment tends to 0 uniformly in $k \le K_n$ as $n \to \infty$.

Proof of Theorem 4

Based on condition (C6), we know that there is a unique estimate $\hat{\theta}$ solving the equation

$$H(\mathbf{Y}, \theta) = n^{-1}\sum_{i=1}^n h_i(Y_i, \theta) = 0.$$

For the MCMB algorithm of Section 2.1 with $a_i = 1$, we have

$$H_j\big(\mathbf{Y}, \hat{\theta}_1^{(k)}, \ldots, \hat{\theta}_{j-1}^{(k)}, \hat{\theta}_j^{(k)}, \hat{\theta}_{j+1}^{(k-1)}, \ldots, \hat{\theta}_p^{(k-1)}\big) = H_j^{*(k)} \tag{A.15}$$

for $j = 1, \ldots, p$ and $k = 1, 2, \ldots$, where H_j is the jth component of $H(\mathbf{Y}, \theta)$ and $H_j^{*(k)} = n^{-1}\sum_{i=1}^n z_{ij}^{*(k)}$, with $z_{ij}^{*(k)}$ as defined in Section 2.1.

Let $H_{jl}(\mathbf{Y}, \theta) = \partial H_j(\mathbf{Y}, \theta)/\partial\theta_l$. With condition (C7), we can rewrite (A.15) as

$$\begin{aligned} &H_{j1}(\mathbf{Y}, \hat{\theta})n^{1/2}\big(\hat{\theta}_1^{(k)} - \hat{\theta}_1\big) + \cdots + H_{jj}\big(\mathbf{Y}, \hat{\theta}\big)n^{1/2}\big(\hat{\theta}_j^{(k)} - \hat{\theta}_j\big) \\ &\quad + H_{j,j+1}(\mathbf{Y}, \hat{\theta})n^{1/2}\big(\hat{\theta}_{j+1}^{(k-1)} - \hat{\theta}_{j+1}\big) + \cdots \\ &\quad + H_{j,p}(\mathbf{Y}, \hat{\theta})n^{1/2}\big(\hat{\theta}_p^{(k-1)} - \hat{\theta}_p\big) = n^{1/2}H_j^{*(k)} + R_{n,k}^{(j)} \end{aligned} \tag{A.16}$$

for $j = 1, \ldots, p$ and $k = 1, 2, \ldots$, where $R_{n,k}^{(j)}$ is the remainder term. From conditions (C6)–(C8), we can show by similar arguments to those in the proofs of Lemma A.3 and Theorem 1 that there exists $K_n \to \infty$ such that $\max_{k\le K_n}|H_j^{*(k)}| = o(1)$, $\theta^{(k)}$ converges to θ uniformly for $k \le K_n$, and $\max_{k\le K_n} E[R_{n,k}^{(j)}]^2 \to 0$ as $n \to \infty$.

For simplicity, we define $\mathbf{Q}_n$ be the matrix whose ijth element is $H_{ij}(\mathbf{Y}, \hat{\theta})$. Let $\mathbf{A}_n = \mathcal{L}(\mathbf{Q}_n)$ and $\mathbf{B}_n = \mathbf{Q}_n - \mathbf{A}_n$. It follows from (A.16) that

$$n^{1/2}(\hat{\theta}^{(k)} - \hat{\theta}) = -\mathbf{A}_n^{-1}\mathbf{B}_n n^{1/2}(\hat{\theta}^{(k-1)} - \hat{\theta}) + \mathbf{A}_n^{-1} n^{1/2} H^{*(k)} + R_{n,k}, \tag{A.17}$$

where $H^{*(k)} = (H_1^{*(k)}, \ldots, H_p^{*(k)})'$ and $\max_{k<K_n} E\|R_{n,k}\|^2 \to 0$ $n \to \infty$.

From the fact that $E_\theta(\partial h_i(Y_i, \theta)/\partial\theta) = -E_\theta h_i(Y_i, \theta)h_i'(Y_i, \theta)$ for the likelihood score, we have $\mathrm{var}(n^{1/2}H^{*(k)}) - \mathrm{diag}(\mathbf{Q}_n) \to 0$. The rest of the proof is then parallel to that of Theorem 1.

Proof of Corollary 1

Note that $E_\beta Y_i = b^{(1)}(x_i'\beta)$, $\mathrm{var}_\beta Y_i = b^{(2)}(x_i'\beta)a(\phi)$, $h_i(Y_i, \beta) = x_i(Y_i - b^{(1)}(x_i'\beta))/a(\phi)$, and $\partial h_i(Y_i, \beta)/\partial\beta = -b^{(2)}(x_i'\beta)x_i x_i'/a(\phi)$. Condition (C7) specializes to (C9) in our model. The second convergence of (C8) follows directly from (C10). To verify the first convergence of (C8), we apply Corollary to theorem 5.4.1 of Chung (1974, p. 125) to obtain

$$n^{-1}\sum_{i=1}^{n} x_i x_i'(Y_i - b^{(1)}(x_i'\beta))^2 - n^{-1}\sum_{i=1}^{n} x_i x_i' b^{(2)}(x_i'\beta)a(\phi) \to 0 \text{ a.s.}$$

Therefore, they both have the same limit. To see (C6), we just note that $n^{-1}\sum_{i=1}^{n} logf_i(y_i; \beta)$ is a concave function of β with its second-order derivative $-n^{-1}\sum_{i=1}^{n} x_i x_i' b^{(2)}(x_i'\beta)/a(\phi)$, which converges to a negative definite matrix.

[Received July 1999. Revised January 2002.]

REFERENCES

Bai, Z. D., and He, X. (1999), "Asymptotic Distribution of the Maximal Depth Estimators for Regression and Multivariate Location," *The Annals of Statistics*, 27, 1616–1637.

Bai, Z. D., Rao, C. R., and Wu, Y. (1992), "M-Estimation of Multivariate Linear Regression Parameters Under a Convex Discrepancy Function," *Statistica Sinica*, 2, 237–254.

Chung, K. L. (1974), *A Course in Probability Theory* (2nd ed.), New York: Academic Press.

Diciccio, T. J., and Efron, B. (1996), "Bootstrap Confidence Intervals" (with discussion), *Statistical Science*, 11, 189–228.

Diciccio, T. J., and Romano, J. P. (1988), "A Review of Bootstrap Confidence Intervals," *Journal of the Royal Statistical Society*, Ser. B, 50, 338–354.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1–26.

Efron, B., and Tibshirani, R. J. (1993), *An Introduction to Bootstrap*, New York: Chapman and Hall.

Freedman, D. A. (1981), "Bootstrapping Regression Models," *The Annals of Statistics*, 9, 1228–1228.

Godambe, V. P., and Kale, B. K. (1991), "Estimating Functions: An Overview," *Estimating Functions* (V. P. Godambe, ed.), Oxford University Press, Oxford, pp. 3–20.

Gutenbrunner, C., Jurečková, J., Koenker, R., and Portnoy, S. (1993), "Tests of Linear Hypotheses Based on Regression Rank Scores," *Journal of Nonparametric Statistics*, 2, 307–331.

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, New York: Springer-Verlag. Berlin.

Hall, P., and Sheather, S. J. (1988), "On the Distribution of a Studentized Quantile," *Journal of the Royal Statistical Society*, Ser. B, 50, 381–391.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, New York: Wiley.

Hannan, E. J. (1970), *Multiple Time Series*. New York: Wiley.

He, X., and Shao, Q. M. (1996), "A General Bahadur Representation of M-Estimators and Its Application to Linear Regression With Nonstochastic Design," *The Annals of Statistics*, 24, 2608–2630.

Hosmer, D. W., and Lemeshow, S. (1989), *Applied Logistic Regression*, New York: Wiley.

Hu, F., and Kalbfleisch, J. D. (1997), "Estimating Equations and the Bootstrap," in *Selected Proceedings of the Symposium on Estimating Equations*, eds. I. V. Basawa, V. P. Godambe and R. L. Taylor, Hayward, CA: IMS, pp. 405–416.

——— (2000), "The Estimating Equation Bootstrap" (with discussion), *Canadian Journal of Statistics*, 28, 449–499.

Hu, F., and Zidek, J. V. (1995), "A Bootstrap Based on the Estimating Equations of the Linear Model," *Biometrika*, 82, 263–275.

Koenker, R. (1994), "Confidence Intervals for Regression Quantiles," in *Asymptotic Statistics: Proceedings of the 5th Prague Symposium on Asymptotic Statistics*, eds. P. Mandl and M. Huskova, Heidleberg: Physica-Verlag.

Lahiri, S. N. (1992), "Bootstrapping M-Estimators of a Multiple Linear Regression Parameter," *The Annals of Statistics*, 20, 1548–1570.

Liu, R. Y. (1988), "Bootstrap Procedures Under Some Non-iid Models," *The Annals of Statistics*, 16, 1696–1708.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.

Parzen, M. I., Wei, L. J., and Ying, Z. (1994), "A Resampling Method Based on Pivotal Estimating Functions," *Biometrika*, 81, 341–350.

Petrov, V. V. (1995), *Limit Theorems of Probability Theory—Sequences of Independent Random Variables*, Oxford, U.K.: Oxford Science Publications.

Shao, J., and Tu, D. (1995), *The Jackknife and Bootstrap*, New York: Springer-Verlag. Berlin.

Searle, S. R. (1982), *Matrix Algebra Useful for Statistics*, New York: Wiley.

Yohai, V. J., and Maronna, R. A. (1979), "Asymptotic Behavior of M-Estimators for the Linear Model," *The Annals of Statistics*, 7, 258–268.

Zhou, Q., and Portnoy, S. L. (1996), "Direct Use of Regression Quantiles to Construct Confidence Sets in Linear Models," *The Annals of Statistics*, 24, 287–306.